

Optimal Mutation Rates on Static Fitness Landscapes

Martin Nilsson

Institute of Theoretical Physics, Chalmers University of Technology and Göteborg University, S-412 96 Göteborg, Sweden

martin@fy.chalmers.se

(February 2, 2008)

We study the evolution of mutation rates for an asexual population living on a static fitness landscape, consisting of multiple peaks forming an evolutionary staircase. The optimal mutation rate is found by maximizing the diffusion towards higher fitness. Surprisingly the optimal genomic copying fidelity is given by $Q_{opt} = e^{-\frac{1}{\ln \nu}}$ (where ν is the genome length), independent of all other parameters in the model. Simulations confirm this theoretical result. We also discuss the relation between the optimal mutation rate on static and dynamic fitness landscapes.

Evolution on the molecular level can be viewed as a diffusion process. The equations describing the time dynamics of a population of gene sequences are a set of discrete diffusion equations with an exponential growth term. The diffusion stems from inaccurate copying of the genome during replication. This enables the population to explore the sequence space, i.e., the space spanned by all possible gene sequences. Point mutations makes the Hamming distance a natural metric on sequence space, which becomes topologically equivalent to a hyper-cube of dimension ν , where ν is the genome length. The high dimensionality makes analysis of the general diffusion process difficult. In this paper we focus on the evolution through a specified path in the hypercube and disregard the dynamics of all other gene sequences. This gives a one dimensional sequence space. We are interested in the optimal mutation rate, which is defined as the mutation rate that maximizes the diffusion speed.

The genome codes mainly for proteins which regulate the chemical reactions within the cell. One of the processes that are under genomic control is the replication of the genome itself. When the genetic material is copied there are replicase enzymes involved. This is important since an unguided base pairing process is highly inaccurate. The enzymes are determined by the genome and the mutation rate of the organism is therefore under evolutionary control. This implies that the mutation rates observed in living organisms have been selected for by Darwinian evolution.

Naively one may think that since most mutations that affect the fitness are deleterious, organisms should evolve as low mutation rates as possible. Measurements of mutation rates however show that organisms have copying fidelities much below what could be expected from this assumption [1,2]. They also show that the genomic mutation rate, i.e., the probability of one or more mutations to occur during one replication of the whole genome, is approximately constant within similar groups of organisms. This is surprising since the copying of the genetic material is a local process and it is the mutation rate per base pair that are directly affected by the replicase enzymes. Most attempts to find an evolutionary explanation for

the observed mutation rates have been based on populations evolving in a changing environment, see e.g., [3–10]. It is easy to understand that a non-zero mutation rate is selected for on a dynamic fitness landscape, since perfect copying will unable adaption to new conditions. Recently a theoretical study has shown that the optimal genomic copying fidelity in a dynamic environment is approximately independent of genome length [10]. The theory also predicts mutation rates of the same order of magnitude as observed for simple DNA based organisms. In this paper we study a different model. The population lives in a static environment, but starts far from the global fitness maximum. A non-zero mutation rate is selected for by maximizing the rate of evolution towards better fit genotypes.

Consider an asexual haploid population of individuals, represented by genomes of length ν . The fitness landscape consists of a number of peaks with superior fitness surrounded by a background. The evolution on this landscape is driven by mutations enabling jumps from one fitness peak to a higher peak in the close neighborhood. We study a population of N gene sequences starting at a low fitness peak which then mutate onto successive fitness peaks of increasing height ($\sigma_1 < \sigma_2 < \dots$). Furthermore we assume the copying fidelity per base, q , to be constant over the genome. The probability of a gene sequence to copy onto itself during one replication event, the genomic copying fidelity, is then given by $Q = q^\nu$. We also assume the probability of an individual on peak σ_{i-1} to produce an offspring on peak σ_i during a replication event to be $p_i(1 - q)^{\alpha_i} q^{\nu - \alpha_i}$. This means that the number of bases where the sequences defining peak σ_{i-1} and σ_i differ is α_i . The factor p_i is an arbitrary combinatorial factor, accounting for possible redundancies in sequence space, alphabet size, etc. All higher fitness peaks, σ_k for $k > i$, are assumed to be further away so that mutations from peak σ_{i-1} can be neglected. The evolution of the relative concentrations x_n is described by differential equation

$$\dot{x}_1 = W_{1,1}\theta_N(x_1) + W_{1,2}\theta_N(x_2) - fx_1$$

$$\dot{x}_2 = W_{2,1}\theta_N(x_1) + W_{2,2}\theta_N(x_2) + W_{2,3}\theta_N(x_3) - fx_2$$

⋮

$$\dot{x}_n = W_{n,n-1}\theta_N(x_{n-1}) + W_{n,n}\theta_N(x_n) + W_{n,n+1}\theta_N(x_{n+1}) - fx_n \quad (1)$$

where the function θ_N is defined as $\theta_N(x_n) = x_n$ if $x > \frac{1}{N}$ and 0 otherwise, and therefore accounts for the limited population size. The factor $f = \sum_i (q^\nu + p_i(1-q)^{\alpha_i})\sigma_i\theta_N(x_i)$ ensures x_i to be normalized as relative concentrations. The matrix elements of W are given by

$$\begin{aligned} W_{n,n} &= q^\nu\sigma_n \\ W_{n-1,n} &= p_n(1-q)^{\alpha_n}q^{\nu-\alpha_n}\sigma_{n-1} \\ W_{n,n+1} &= p_{n+1}(1-q)^{\alpha_{n+1}}q^{\nu-\alpha_{n+1}}\sigma_{n+1} \end{aligned} \quad (2)$$

We start with a population that consists of individuals on the first peak σ_1 , i.e., we define the initial values as

$$x_i(0) = \begin{cases} 1 & i = 1 \\ 0 & i \neq 1 \end{cases} \quad (3)$$

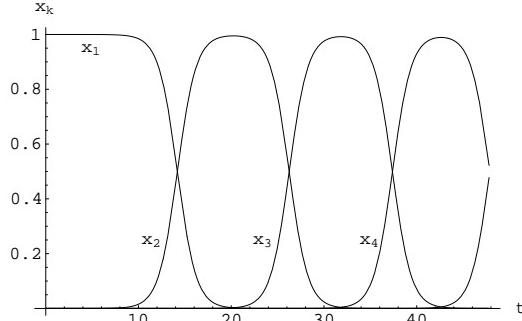


FIG. 1. The time dynamics of Eq. 1 is simulated numerically. When the population diffuses off the initial peak σ_1 it starts evolving to peaks with higher and higher fitness. The parameters used in this plot are $\nu = 100$, $\sigma_i = i$, $p = 0.01$, $Q = 0.99$ and $N = 10^6$.

The infinite population size limit of Eq. 1 corresponds to a discrete normalized one-dimensional diffusion equation with an exponential growth term. However, this limit is not interesting for realistic systems since it does not allow propagating distributions of concentrations localized in sequence space. If the fitness grows faster than linearly for example, the concentration on fitness peaks far from the starting point grow large *before* the concentrations on peaks closer to the origin. This bizarre effect stems from the exponential growth of very small (exponentially decaying with the distance from the origin) but non-zero concentrations over all the fitness peaks shortly after the start.

In this model we implicitly assume the mutation rates to evolve much slower than the fitness, i.e. there are no significant changes in the mutation rate during the evolution from one fitness peak to the next peak.

The optimal copying fidelity q_{opt} is defined by maximizing the diffusion speed towards genotypes with superior fitness. Mathematically this corresponds to minimizing the time T it takes for the concentration x_n on peak σ_n to reach its maximum, when the population starts at the proceeding peak σ_{n-1} . At the time when mutants from peak σ_{n-1} have enabled the concentration x_n to become large enough, i.e. $x_n > \frac{1}{N}$, exponential growth will start with initial concentration proportional to $p_n(1-q)^{\alpha_n}$. Since the population at this time is localized around peak n the concentration x_n is described approximately by

$$x_n(t) \sim \frac{\gamma e^{q^\nu \sigma_n t}}{e^{q^\nu \sigma_{n-1} t} + \gamma e^{q^\nu \sigma_n t} + \gamma^2 e^{q^\nu \sigma_{n+1} t}} \quad (4)$$

where $\gamma = p_n(1-q)^{\alpha_n}$. The denominator normalizes x_n by summing the absolute growth in the surrounding of peak n , see Fig. 2. The time T when $x_n(t)$ has a maximum can be found by solving $\frac{dx_n(t)}{dt} = 0$, giving

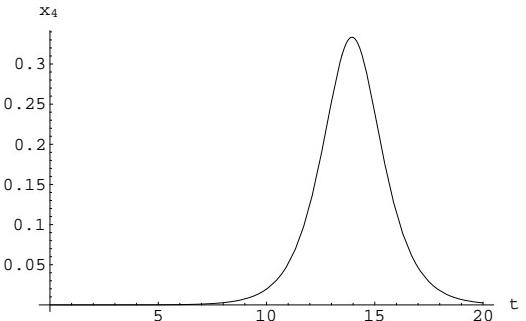


FIG. 2. The picture shows x_4 given by Eq. 4. The relevant parameter are the same as in Fig. 1; $\nu = 100$, $\sigma_i = i$, $p = 0.01$ and $Q = 0.99$. The maximum occurs at time $T \approx 14$ (defined as the time from the last peak's maximum). This is in agreement with the numerical solutions shown in Fig. 1.

$$T = -\frac{1}{\sigma_{n+1} - \sigma_{n-1}} \cdot \frac{\ln(\gamma^2 \kappa)}{q^\nu} \quad (5)$$

where $\kappa = \frac{\sigma_{n+1} - \sigma_n}{\sigma_n - \sigma_{n-1}}$. The diffusion speed is defined as $v = \frac{1}{T}$. By making the approximation $\kappa \approx 1$, we can write

$$V = -\frac{\sigma_{n+1} - \sigma_{n-1}}{2} \cdot \frac{q^\nu}{\ln(\gamma)} \quad (6)$$

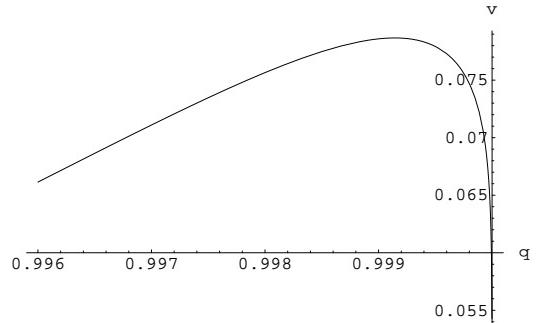


FIG. 3. The figure shows $V(q)$ given by Eq. 6. The maximum gives the optimal copying fidelity q_{opt} . Parameters used in the figure are $\nu = 100$, $\sigma_i = i$ and $p = 0.01$. The shape of the curve is not sensitive to the parameter values, as long as $\nu \gg 1$.

The optimal copying fidelity q_{opt} is defined to maximize the diffusion speed, and can therefore be derived by finding the maximum of $V(q)$ in Eq. 6 (see Fig. 3). Setting the derivative to zero, $\frac{dV}{dq} = 0$, and noting that $q \approx 1$ gives the equation:

$$1 + \frac{1}{\nu(1-q)(\ln(\bar{p}) + \ln(1-q))} = 0 \quad (7)$$

We are interested in the limit where the genome length is large. In this limit the first term in the denominator (involving \bar{p}) can be neglected. Eq. 7 then reduces to

$$\nu(1-q)\ln(1-q) = -1 \quad (8)$$

There is no closed analytic expressions for the solution to this equation, but a converging iterative expression can be found for the optimal copying fidelities

$$q_{opt} = 1 - \frac{1}{\nu \ln(\nu \ln(\nu \ln(\dots)))} \\ Q_{opt} \approx e^{-\frac{1}{\ln \nu}} \quad (9)$$

It is surprising that the optimal genomic copying fidelity depends so weakly on the genome length, and even more surprising that it is independent of all other parameters in the model. This independence is both interesting and important, especially since we start by assuming a specific path for evolution. As it turns out the optimal mutation rate does not depend on the particular path chosen. The insensitivity of Q_{opt} when the genome length varies can be seen by considering biologically plausible genome lengths, see Fig. 5. Note that the genomic copying fidelity increases with genome length.

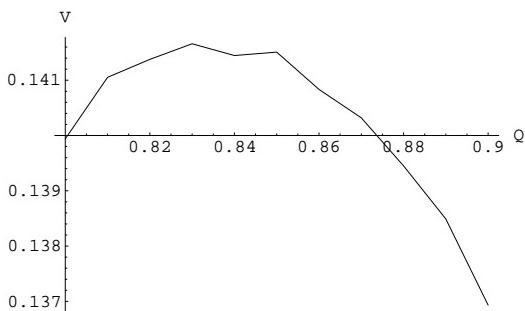


FIG. 4. The figure shows the region where $V(q)$ has a maximum, calculated by numerical simulations of Eq. 1. Parameter settings in the simulations were $p = 0.01$, $\alpha = 1$, $N = 10^8$ and $\nu = 1000$. The minimum occurs approximately at the point predicted by Eq. 9, i.e., $Q_{opt} = 0.86$.

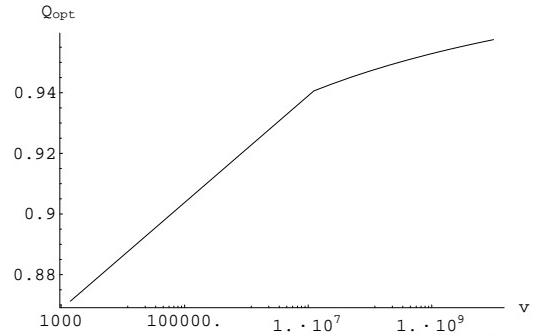


FIG. 5. The figure demonstrates how weakly Q_{opt} scales with genome length.

In simulations of a population consisting of 2000 individuals with genome length $\nu = 70$ on a rugged fitness landscape (created by an elementary folding algorithm for calculating secondary structures of gene sequences), Fontana and Schuster [11] find that the rate of evolution is maximal approximately at $\mu = 0.003$. This is in close agreement with the mutation rate as predicted by Eq. 9 for genome length 70, $\mu_{opt} = 0.0025$.

The optimal copying fidelity given in Eq. 9 can also be derived using a more intuitive argument. The argument also shows more clearly how evolved mutation rates on static fitness landscapes relate to evolved mutation rates in dynamic environments. The rate of growth between two peaks, with fitness difference $\Delta\sigma$, is given by $e^{Q\Delta\sigma t}$. The diffusion from an occupied peak to the next is proportional to $(1-q)^\alpha$, where α measures the distance in sequence space between the peaks. The time, T , it takes for a population to evolve from one peak to another will therefore be given by the solution of the equation $(1-q)^\alpha e^{Q\Delta\sigma t} = 1$, i.e. $T \sim -\frac{\alpha \ln(1-q)}{q^\nu \Delta\sigma}$. Organisms, free to change their mutation rates, evolve a copying fidelity q_{opt} that minimizes $T(q)$. Deriving an expression for the equation $\frac{dT}{dq} = 0$, using $q^\nu \approx 1$, gives $\frac{1}{1-q} + \nu \ln(1-q) = 0$, which is equivalent to Eq. 8 and is solved by $Q_{opt} \approx e^{-\frac{1}{\ln \nu}}$.

In a recent paper [10], the evolution of mutation rates on a dynamic fitness landscape was studied. The fitness landscape consists of a single peak moving around in sequence space, shifting position on average once every τ generations. The relative selective advantage for a sequence on the fitness peak is σ . A shift of the peak consist of α changes of bases in the sequence defining the fitness peak. Since an individual in the population needs to produce offspring that are able to follow the shifts of the fitness peak, a non-zero mutation rate is selected for. It turns out that finding the optimal copying fidelity is equivalent to minimizing $(1-q)^\alpha e^{Q\sigma\tau}$ with respect to q . This is the same expression as for the growth rate between fitness peaks on a static landscape. However, in the dynamic case the growth over a cycle, consisting of a shift and a static period, is be optimized rather than the time to evolve from one peak to the next. More gen-

erally, if the evolution of mutation rate is driven by a dynamic environment it will be selected to optimize the growth on the changing fitness landscape, whereas on a static landscape the mutation rate maximizing the rate of evolution towards higher fit genotypes will be selected for. Maximizing the growth on a dynamic landscape gives $Q_{dyn} = e^{-\frac{\sigma}{\sigma\tau}}$.

There are some fundamental differences between the two models presented above. In the model based on dynamic fitness landscapes the population dynamics is driven by external changes of the environment. The organisms have to passively wait for the environment to change and then adapt to the new fitness landscape. In the model based on rugged fitness landscapes the situation is different. There always exist a higher fitness peak in the close neighborhood so the population has to minimize the time for diffusing to and growing large on the higher peak. Hence the population should actively search the surroundings in sequence space. The main difference between the models is therefore the preexistent of higher fitness peaks close in sequence space, which results in very different optimal mutation rates.

The genomic copying fidelity in both the static and dynamic case is approximately independent of genome length, a phenomenon that is also observed in nature. To be more precise, experiments show that the genomic copying fidelity is approximately constant within groups of similar organisms, e.g., simple DNA-based organisms have $Q \approx 0.996$ whereas RNA based retroviruses have $Q \approx 0.9$ [1]. Simple DNA based organisms for example have much too low mutation rates to be explained by evolution on the static landscapes studied in this paper. Retroviruses on the other hand show mutation rates that are in agreement with the predictions made in this paper. However, they may also be explained by mutation rates evolved as a response to a changing environment as discussed above. It is therefore unclear whether the major force behind the evolution of mutation rates for retrovirus is maximizing the evolution rate towards higher fitness or maximizing the growth in a changing environment (caused by the immune system).

In conclusion, we show that the optimal genomic copying fidelity, i.e., that which optimizes the rate of evolution, on a rugged fitness landscape can be written as $Q_{opt} = e^{-\frac{1}{\ln(\nu)}}$, where ν is the genome length. The optimal genomic copying fidelity on rugged fitness landscapes is predicted to be around 0.9 for realistic genome lengths ($\nu \in [10^3, 10^{10}]$). Of the mutation rates observed

in nature, retroviruses (including HIV) confirm this prediction. The model presented here therefore presents a possible explanation for the observed mutation rates for retro viruses.

The author would like to thank Jennie Jacobi and “Mullbäret” for providing a nice and stimulating environment while working on the ideas behind this paper. Thanks are also due to Mats Nordahl and Johan Ivarsson for valuable comments on the manuscript.

-
- [1] J.W. Drake, B. Charlesworth, D. Charlesworth, and J.F. Crow. Rates of spontaneous mutation. *Genetics*, 148(4):1667–86, 1998.
 - [2] J.W. Drake. Rates of spontaneous mutation among RNA viruses. *Proc. Natl. Acad. Sci. U.S.A.*, 90(9):4171–5, 1993.
 - [3] M Kimura. On the evolutionary adjustment of spontaneous mutation rates. *Genet. Res.*, 9:23–24, 1967.
 - [4] E.G. Leigh. The evolution of mutation rates. *Genetics Suppl.*, 73:1–18, 1973.
 - [5] K. Ishii, H. Matsuda, Y. Iwasa, and A. Saskai. Evolutionarily stable mutation rate in a periodically changing environment. *Genetics*, 121:163–174, 1989.
 - [6] J.H. Gillespie. Mutation modification in a random environment. *Evolution*, 35:468–476, 1981.
 - [7] S.P. Otto and Y. Michalakis. The evolution of recombination in changing environments. *Trends Ecol. Evol.*, 13(4):145–151, 1998.
 - [8] S.P. Otto and M.W. Feldman. Deleterious mutations, variable epistatic interactions, and the evolution of recombination. *Theor. Popul. Biol.*, 51(2):134–147, 1997.
 - [9] G. Ochoa, I. Harvey, and H. Buxton. On recombination and optimal mutation rates. In *Proceedings of Genetic and Evolutionary Computation Conference (GECCO-99)*, Orlando, Florida, USA, July 1999.
 - [10] M. Nilsson and N. Snoad. Optimal Mutation Rates on Dynamic Fitness Landscapes. LANL e-print archive: physics/0004042, 2000.
 - [11] W. Fontana and P. Schuster. A computer model of evolutionary optimization. *Biophysical Chemistry*, 26:123–147, 1987.
 - [12] L. Tsimring, H. Levine and D. Kessler. RNA Virus Evolution via a Fitness-Space Model. *Phys. Rev. Lett.*, 76(23), 4440–4443, 1996.
 - [13] D. Kessler and H. Levine. Mutator Model on a Smooth Evolutionary Landscape. *Phys. Rev. Lett.*, 80(9), 2012–2015, 1998.